Probabilistic forecast reconciliation with applications to wind power and electric load

Jooyoung Jeon[a,b,*], Anastasios Panagiotelis[c], Fotios Petropoulos[a]

[a] *School of Management, University of Bath*
[b] *Graduate School of Engineering Practice, Seoul National University*
[c] *Department of Econometrics & Business Statistics, Monash University*

Abstract

New methods are proposed for adjusting probabilistic forecasts to ensure coherence with the aggregation constraints inherent in temporal hierarchies. The different approaches nested within this framework include methods that exploit information at all levels of the hierarchy as well as a novel method based on cross-validation. The methods are evaluated using real data from two wind farms in Crete and electric load in Boston. For these applications, optimal decisions related to grid operations and bidding strategies are based on coherent probabilistic forecasts of energy power. Empirical evidence is also presented showing that probabilistic forecast reconciliation improves the accuracy of the probabilistic forecasts.

Keywords: Forecasting, Temporal hierarchies, Cross-validation, Aggregation, Renewable energy generation

1. Introduction

Data are often arranged in hierarchies characterised by an aggregation structure that holds for all realised values; for example, the annual sum of monthly data series will be equivalent to annual data series. When forecasts are independently produced for different series or levels within a hierarchy these aggregation constraints will not hold, a property known as *incoherence*. To ensure that operational decisions are aligned, a rich literature has emerged on forecast reconciliation (Athanasopoulos, Ahmed, and Hyndman, 2009; Hyndman et al., 2011; Athanasopoulos et al., 2017; Wickramasuriya, Athanasopoulos, and Hyndman, 2018). These approaches not only ensure

*Correspondence: Jooyoung Jeon
Email addresses: j.jeon@bath.ac.uk (Jooyoung Jeon), Anastasios.Panagiotelis@monash.edu (Anastasios Panagiotelis), f.petropoulos@bath.ac.uk (Fotios Petropoulos)

that forecasts are coherent but also lead to improvements in forecast accuracy by combining information and forecasts from different hierarchical levels. However, a shortcoming of these approaches is their focus on point forecasting despite the increasing importance of probabilistic forecasts on decision-making (Gneiting and Katzfuss, 2014).

In a similar way to point forecasts, probabilistic forecasts can be produced independently for each level in the hierarchy. However, independent series cannot be coherent since the aggregation constraint induces dependence between the variables. This paper focuses on the open question of reconciling *probabilistic* forecasts. Our contributions are as follows:

- We propose approaches for reconciling probabilistic forecasts that ensure coherence. We refer to this as 'probabilistic forecast reconciliation' since information is combined from density forecasts at all hierarchical levels.
- We focus on producing coherent probabilistic forecasts in the temporal rather than in the cross-sectional hierarchical setting, although we note that the proposed approaches are general enough to handle both settings.
- We propose an approach that considers reconciliation weights via a cross-validation procedure. This is the first time that a cross-validation procedure is used to produce-coherent forecasts in either the point or probabilistic domain.

To the best of our knowledge, the only other paper to tackle the issue of coherent probabilistic forecasts is that of Ben Taieb, Taylor, and Hyndman (2017) where, with the exception of the mean and variance, the construction of a coherent probabilistic forecast relies on a bottom-up approach. In particular, Ben Taieb, Taylor, and Hyndman (2017) construct a coherent probabilistic forecast in a bottom-up fashion where the dependency between nodes at each level is modelled by reordering quantile forecasts as suggested by Arbenz, Hummel, and Mainik (2012). The method we propose is distinct from Ben Taieb, Taylor, and Hyndman (2017) in two ways. First, our proposed method is a true reconciliation method, since each probabilistic forecast is based on information from the entire density and for all nodes in the hierarchy. Second, our problem focuses on temporal aggregation of density forecasts which provides a distinct case since dependence within each level can be obtained directly rather than through copula modelling.

The methods we propose are evaluated using wind power and electric load data measured at various frequencies ranging from hourly to daily. These applications are chosen for two main reasons.

First, due to the highly volatile nature of wind power generation and electric load, informed decision-making depends not only on point forecasts but on probabilistic considerations. For instance, dispatch and risk management decisions on unit commitment may be based on the probability that a wind farm supplies at least $300kWh$ between midnight and 6am the following day. Second, wind farm operators, grid system operators and electricity traders are each required to make decisions based on different forecast horizons and sampling frequencies. Third, accurate probabilistic forecast in lead times and resolutions typically up to 24 hours ahead facilitates optimal scheduling of energy storage and peer to peer (P2P) energy trading under uncertainties between prosumers (Morstyn et al., 2018). As such coherent probabilistic forecasts are crucial to ensure aligned decision-making. Our empirical results demonstrate that the proposed reconciliation methods improve the accuracy of probabilistic forecasts, with more substantial improvements at higher aggregation levels.

In the next section, we review the literature on forecast reconciliation for point forecasting. Section 3 presents our newly proposed methods for producing coherent, reconciled density forecasts. Section 4 introduces our dataset and describes the approach used to obtain base forecasts that can subsequently be reconciled. Section 5 describes the empirical results of the various reconciliation methods considered in Section 3. The final section provides a summary and conclusion.

2. Background: Point Forecast Reconciliation

2.1. Cross-sectional Hierarchical Reconciliation of Point Forecasts

Data are often organised in hierarchical aggregation structures. For example, a company may organise its five stock keeping units (SKUs) into two categories, as depicted in Figure 1. If the historical data at the bottom level (SKU) are available, then data at every other level can be calculated using appropriate aggregations. Forecasts may be produced at any of the three levels of the hierarchy. However, if forecasts are independently produced at all levels they will not be coherent. For example, the sum of the forecasts of SKUs 1, 2 and 3 in Figure 1 is not guaranteed to be the same as the forecast of Category 1.

One way to tackle this issue is to simply produce forecasts on a single hierarchical level. For example, forecasts can be produced only on the very bottom level, and then aggregated to the higher levels in the hierarchical structure, an approach known as the *bottom-up* approach (see for example Dangerfield and Morris, 1992; Zellner and Tobias, 2000; Athanasopoulos, Ahmed, and Hyndman,

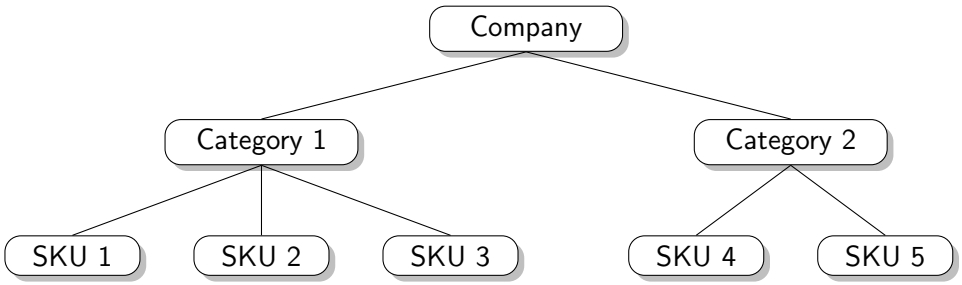

Figure 1: A cross-sectional hierarchy.

2009). In some cases, the bottom-level data may be too granular or noisy, rendering the forecasting task difficult. Alternatively, forecasts may be produced at the very top-level and then appropriately disaggregated to lower level forecasts, an approach known as *top-down* (Lütkepohl, 1984; Fliedner, 1999; Gross and Sohl, 1990). Disaggregation of the forecasts to lower levels may be based on historical or predicted proportions of the lower level data (Athanasopoulos, Ahmed, and Hyndman, 2009). The top-down approach has the disadvantage of information loss, as aggregated series may not reflect the individual characteristics of their descendants. Finally, forecasts can also be produced at a middle level; forecasts for higher/lower levels nodes can be calculated by appropriate aggregation/disaggregation of the middle-level forecasts. This approach is known as *middle-out*, a conceptual combination of the bottom-up and top-down approaches.

A shortcoming of the methods above is that forecasts are only based on information at a single level of the hierarchy. The optimal combination method (Athanasopoulos, Ahmed, and Hyndman, 2009; Hyndman et al., 2011) overcomes this problem by tackling hierarchical forecasting in two stages. In the first stage, point forecasts are produced for all series at all levels independently. These first stage forecasts are referred to as 'base' forecasts. In the second stage, these forecasts are adjusted or 'reconciled' to ensure coherence with aggregation constraints. More specifically the reconciled forecast for each node is formed as a weighted combination of the 'base' forecasts of all nodes, in a way that ensures coherence for the hierarchy overall. The key advantage of reconciliation is that information is used at all levels of the hierarchy in contrast to the approaches described in the previous paragraph that focus on a single level. More recently, Hyndman, Lee, and Wang (2016) propose algorithms for fast computation of coherent hierarchical forecasts, and Wickramasuriya, Athanasopoulos, and Hyndman (2018) suggest calculating coherent forecasts through trace minimisation.

2.2. Temporal Hierarchical Reconciliation of Point Forecasts

A time series can be aggregated or disaggregated to create alternative frequency (or resolution) as needed. Time series at different frequencies will exhibit different characteristics. Seasonality and noise will be amplified in lower aggregation levels (higher frequencies), while the long-term trend can be more easily estimated using higher aggregation levels (lower frequencies) (Kourentzes, Petropoulos, and Trapero, 2014; Spithourakis et al., 2014). Similar to the case of cross-sectional aggregation, forecasts produced using data at different frequencies will not generally be coherent. For example, the sum of the forecasts for the next three months, produced using data measured at the monthly frequency, will not equal to the one-step-ahead quarterly forecast based on data measured at a quarterly frequency. This problem is particularly relevant for aligning decisions across the different departments within a company (operations, sales, finance, marketing, strategy), which usually operate at different data frequencies.

Similarly to cross-sectional aggregation, the issue of incoherent forecasts at different temporal aggregation levels can be addressed either by combining (reconciling) the forecasts from multiple aggregation levels or by producing forecasts for a single temporal aggregation level and then deriving the forecasts at the other levels as discussed previously.

Nikolopoulos et al. (2011) show empirically that in the context of intermittent demand there exists an optimal aggregation level, unique to each series, and proposed the Aggregate-Disaggregate Intermittent Demand Approach (ADIDA), where forecasts are produced at a (single) higher aggregation level and the lower level forecast is subsequently produced by disaggregation. This approach is particularly relevant for slow moving data, as temporal aggregation will result in series with a lower degree of intermittence (Petropoulos, Kourentzes, and Nikolopoulos, 2016). Rostami-Tabar et al. (2013) derive analytical results that imply that improvement in forecasting performance is a function of the aggregation level, under specific data generating processes.

The idea of using aggregation/disaggregation for forecasting was further extended to derive the combined forecasts from forecasts simultaneously produced at multiple temporal aggregation (MTA) levels by Kourentzes, Petropoulos, and Trapero (2014) and Petropoulos and Kourentzes (2014). MTA was also applied to the context of intermittent demand (Petropoulos and Kourentzes, 2015), and Kourentzes and Petropoulos (2016) propose an extension to incorporate the effects of external variables. More recently, Athanasopoulos et al. (2017) express the MTA approach as a

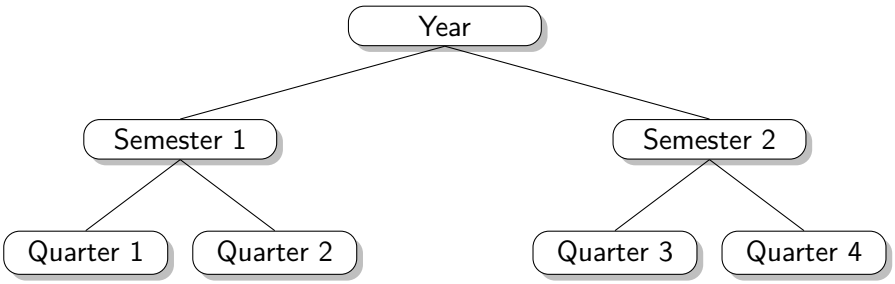

Figure 2: A temporal hierarchy.

hierarchical concept using a temporal hierarchy for forecasting. A simple temporal hierarchy is depicted in Figure 2, where the bottom-level data are at a quarterly frequency (1 quarter per node), middle-level data are at a semesterly frequency (2 quarters per node), and the top-level represents the yearly frequency (4 quarters for the top-level node).

The representation of multiple temporal aggregation as temporal hierarchies allows for the application of the approaches designed for cross-sectional hierarchies, such as bottom-up, top-down, middle-out and least squares combination. In the context of temporal hierarchies, Athanasopoulos et al. (2017) consider weighted least squares reconciliation with three choices of weights. These are, in increasing order of complexity, structural scaling (weights given by the number of disaggregate series that are summed together to form an aggregate), series variance scaling (weights given by in-sample variance of errors for each series) and hierarchy variance scaling (weights given by in-sample variance of errors for each node). Athanasopoulos et al. (2017) show empirically that simpler scaling approximations provide better results, especially as the frequency of the bottom level increases.

The applicability of forecast reconciliation methods designed for cross sectional hierarchies to temporal hierarchies is possible since the reconciliation stage is identical for both data structures. However, the typical methods used to obtain base forecasts reflect important differences between cross-sectional and temporal hierarchies. Each node in a cross-sectional hierarchy corresponds to a single time series with all series measured in the same frequency; for example, the hierarchy depicted in Figure 1 consists of 8 series. In practice, producing base forecasts for cross-sectional hierarchies involves selecting and fitting as many models as the number of nodes (alternatively judgemental forecasts can be used for some or all nodes).

In contrast, in a temporal hierarchy, each *level* corresponds to a time series with its own particular frequency. For example, the hierarchy depicted in Figure 2 consists of 3 series, yearly at the

top level, bi-yearly at the middle level and quarterly at the bottom level. In practice, base forecasts are produced by selecting and fitting one model per temporal aggregation level. Obtaining base forecasts for the entire hierarchy involves using different models to produce forecasts at different hierarchical levels. For instance, for the hierarchy of Figure 2, using the model fit at quarterly frequency, forecasts are produced at horizons up to four quarters ahead, using the model fit at bi-yearly frequency, forecasts are produced at horizons up to two semesters ahead while using the model fit at annual frequency, forecasts are produced one-year ahead. In this example, the term 'one-year ahead forecast' can be somewhat ambiguous since it can refer to a whole hierarchy of forecasts some of which are bi-annual and quarterly. In the remainder of the paper, where the context is unclear, we use the term 'cycle' to refer to a realisation or forecast of the entire hierarchy.

3. Probabilistic Forecast Reconciliation

Before introducing the notation, we briefly discuss the details of how temporal hierarchies can be constructed. The highest frequency at which the data are measured as well as the cycle are both usually given by the requirements of the forecasting problem. In this case, we let $\boldsymbol{f}$ be a vector so that f_l is the sampling interval (measured in the highest frequency time units) for level l of the hierarchy. Then, $\boldsymbol{f}$ can be set according to the factors of f_1, i.e. the number of highest frequency time units in a cycle. In our applications in Section 4, the highest frequency at which data are measured is hourly and a cycle is a day so $\boldsymbol{f} = [24, 12, 8, 6, 4, 3, 2, 1]$. We note that partitions of varying interval length within the same level could be used if there are good operational reasons for doing so with no loss of generality with respect to our proposed reconciliation methods. For the remainder of this section we will illustrate our proposed methods using the hierarchy in Figure 2 as an example. Here, data is measured quarterly, the cycle is a year and we set $\boldsymbol{f} = [4, 2, 1]$.

Let x^t_{j,f_l} be the realisation of a variable recorded on cycle t during the j^{th} period of the cycle while sampling at an interval of f_l. For example, $x^1_{1,f_1} = x^1_{1,4}$ denotes demand for the first year (first four quarters), $x^3_{2,2}$ denotes the demand for the second semester of the third year and $x^5_{3,1}$ denotes demand for the third quarter of the fifth year. Let the scaled data be denoted by vectors $\boldsymbol{z}^t_l := (1/f_l)(x^t_{1,\boldsymbol{f}_l}, x^t_{2,\boldsymbol{f}_l}, \ldots)'$ for all $l = 1, \ldots, L$, where L is the number of levels of the hierarchy ($L = 3$ for the example hierarchy in Figure 3). Then, $\boldsymbol{z}^t_l$ is the vector of the realisations of all nodes at level l, scaled to be in the same units as the bottom level L, i.e. the highest resolution. This scaling

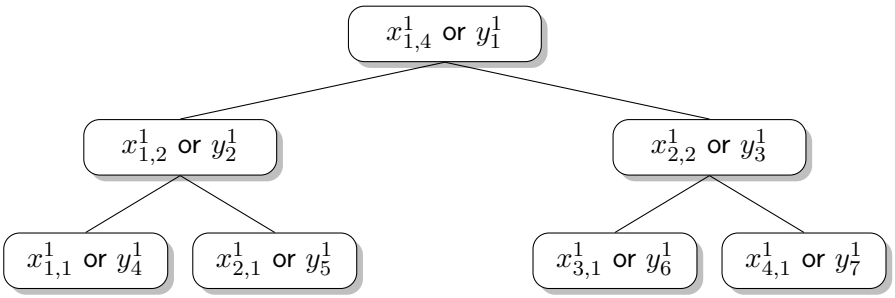

Figure 3: An illustration of notation for a temporal hierarchy.

allows us to avoid the complex scale conversion in the density reconciliation between any levels and to interpret reconciliation as forecast combination between levels. Afterwards, the probabilistic forecasts can be rescaled back to the original units for each level. Finally, let $\boldsymbol{y}^t := (\boldsymbol{z}_1'^t, \ldots, \boldsymbol{z}_L'^t)'$. The notation y_i^t will be used to denote the i^{th} scalar element of $\boldsymbol{y}^t$ for $i = 1, \ldots, M$, where M is the number of nodes in the hierarchy (e.g. $M = 7$ in Figure 3).

3.1. Coherent and Reconciled Point Forecasts

Before describing our new methodologies for probabilistic forecast reconciliation we briefly review the concepts of coherence and forecast reconciliation in the point forecasting setting. By coherence we mean any vector $\boldsymbol{y}^t$ for which the aggregation constraints implied by the hierarchy hold. This can be expressed as $\boldsymbol{y}^t = \boldsymbol{S}\boldsymbol{z}_L^t$. The matrix $\boldsymbol{S}$ is a $M \times m$ matrix that encodes the aggregation constraints and recovers a full set of coherent forecasts from bottom level forecasts, with m equal to the number of bottom level forecasts. For the simple hierarchy in Figure 3, $\boldsymbol{S}$ is given by

$$\boldsymbol{S} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

All realisations of the data are guaranteed to be coherent, i.e. $\boldsymbol{y}^t = \boldsymbol{S}\boldsymbol{z}_L^t$ for all t. Forecasts with this property are called coherent forecasts.

As discussed in Section 2, forecast reconciliation refers to a process by which a vector of incoherent forecasts is made coherent. Letting $\hat{\boldsymbol{y}}$ be a vector of 'base' forecasts, then a reconciled point forecast is given by $\tilde{\boldsymbol{y}} = \boldsymbol{SP}\hat{\boldsymbol{y}}$. The matrix $\boldsymbol{P}$ is a $m \times M$ matrix that forms new point forecasts for the bottom level of the hierarchy as linear combinations of the base point forecasts of all nodes. These are subsequently aggregated up to a full hierarchy by pre-multiplying by $\boldsymbol{S}$. Taken together, $\boldsymbol{SP}$ is a matrix that maps any vector in $\mathbb{R}^M$ to the region where all aggregation constraints must hold.

3.2. Coherent and Reconciled Probabilistic Forecasts

Some care must be taken in extending concepts such as coherent forecasts and reconciled forecasts to the probabilistic setting. Formal definitions of coherent probabilistic forecasts are provided in Ben Taieb, Taylor, and Hyndman (2017) and Gamakumara et al. (2018). In brief, a coherent probabilistic forecast is an M-dimensional multivariate distribution which, due to the degeneracy induced by the aggregation constraints, is only supported on an m-dimensional linear subspace of $\mathbb{R}^M$. To visualise this, consider a trivariate density where $y_{Tot} = y_A + y_B$. In this case all probability should be concentrated in the region where the aggregation constraint holds, which is a 2-dimensional plane within 3-dimensional space. This is depicted in Figure 4 where the red points are generated from a coherent density and therefore all lie on the grey 2D plane spanned by the columns of $\boldsymbol{S}$. For points simulated from a coherent probabilistic forecast, there should be no values of y_{Tot}, y_A and y_B for which the constraint does not hold. All regions in 3D space that do not intersect with a region on the grey 2D plane are assigned zero probability.

Rather than work with analytical expressions for densities, we will instead aim to obtain a sample from the joint predictive distribution of all nodes in the hierarchy. The step that is analogous to obtaining base point forecasts is to generate a sample from the distribution $f\left(\boldsymbol{y}^{t+h}|\mathcal{F}^t\right)$, where $\mathcal{F}^t$ represents all the information up to time t. The key difference with the point forecasting case is that a sample of N vectors from the predictive distribution are produced rather than a single vector of forecasts. Denoting the i^{th} vector from this sample as $\hat{\boldsymbol{y}}_i^{t+h|t}$, we can store these in a matrix as $\hat{\boldsymbol{Y}} = \left(\hat{\boldsymbol{y}}_1^{t+h|t}, \ldots, \hat{\boldsymbol{y}}_N^{t+h|t}\right)$. Typically there is no guarantee that the aggregation constraints will hold for each (or in fact any) of the columns of $\hat{\boldsymbol{Y}}$.

However, if $\hat{\boldsymbol{Y}}$ is pre-multiplied by a suitable matrix to give $\tilde{\boldsymbol{Y}} = \boldsymbol{SP}\hat{\boldsymbol{Y}}$, the columns of the resulting matrix do respect the aggregation constraints. These can therefore be thought of as

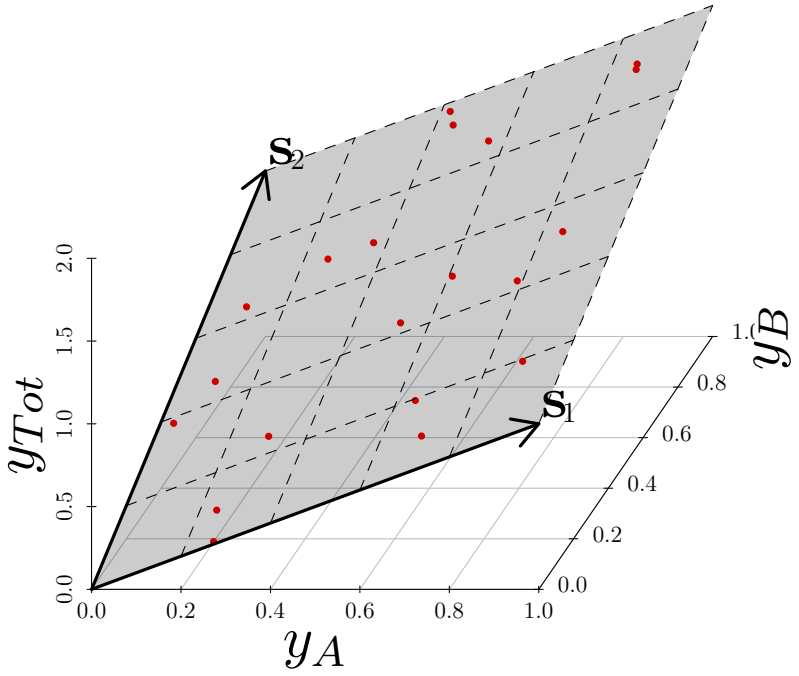

Figure 4: Depiction of a 3 dimensional hierarchy where the red points have been simulated from a coherent probabilistic forecast. All red points lie on a 2D plane (colored grey) spanned by the columns of $\boldsymbol{S}$. For a coherent probabilistic forecast the probability of a point lying in a region that does not include this plane is zero.

observations sampled from the reconciled probabilistic forecast. In this way existing reconciliation methods for the mean are extended to a probabilistic setting. To summarise, the process for forming probabilistic forecasts consists of two stages. In the first stage a sample is obtained from an estimate of the joint density $f\left(\boldsymbol{y}^{t+h}|\mathcal{F}_l^t\right)$, and in the second, each sampled vector is reconciled by pre-multiplying by a matrix that maps each vector to be coherent. At the first stage there are alternative practical approaches to constructing a joint sample, while at the second stage there are alternative reconciliation matrices that can be used. We now discuss each of these stages in detail.

3.3. Construction of Base Probabilistic Forecasts

The first stage of our procedure, namely to obtain a matrix $\hat{\boldsymbol{Y}}$ is itself broken down into two steps. In the first step, each level will be modelled independently. Let $\hat{\boldsymbol{Z}}_l$ be a $(f_1/f_l) \times N$ matrix defined similarly to $\hat{\boldsymbol{Y}}$. Then, its columns are observations sampled from the joint predictive

distribution but only using nodes in level l, i.e. $f\left(\boldsymbol{z}_l^{t+h}|\mathcal{F}_l^t\right)$. A sample from this joint density can be produced by forming multi-step ahead forecasts in the usual recursive fashion and, as a consequence, the dependence within each level is preserved. In the second step, we consider three alternatives for forming a sample $\hat{\boldsymbol{Y}}$ using all $\hat{\boldsymbol{Z}}_l$. Each of these alternatives can be thought of as capturing the dependence between the elements of $\hat{\boldsymbol{Y}}$ in a different way - the appeal of these methods is that they avoid the challenge of modelling dependence explicitly.

3.3.1. Stacked Sample

The most straightforward way to form $\hat{\boldsymbol{Y}}$ is to simply concatenate the matrices $\hat{\boldsymbol{Z}}_l^{t+h|t}$ which we refer to as the 'stacked' sample.

$$\hat{\boldsymbol{Y}^S} = \begin{bmatrix} \hat{\boldsymbol{Z}}_1 \\ \hat{\boldsymbol{Z}}_2 \\ \vdots \\ \hat{\boldsymbol{Z}}_L \end{bmatrix} \tag{2}$$

Using this approach leads to a joint distribution that preserves the dependence within each level but effectively assumes independence between levels, since each $\boldsymbol{Z}_l$ is obtained from an independent modelling process.

3.3.2. Ranked Sample

An alternative to the stacked sample involves ordering the elements in each row of $\hat{\boldsymbol{Y}}^S$ in ascending (or descending) order after concatenation. We refer to this as the 'ranked sample' denoted $\hat{\boldsymbol{Y}}^R$. The rows of $\hat{\boldsymbol{Y}}^R$ will have a comonotonic dependence structure with respect to one another, and this approach can therefore be expected to work well in applications where dependence is high. Furthermore, the i^{th} column of $\hat{\boldsymbol{Y}}^R$ can be thought of as a vector of the $(i/N)^{th}$ quantiles, each element corresponding to a different node. As such, this approach also has an interpretation as a method that reconciles quantiles. It also has similarities to the combination of probabilistic forecasts by Lichtendahl, Grushka-Cockayne, and Winkler (2013). Whereas Lichtendahl, Grushka-Cockayne, and Winkler (2013) focus on combining probabilistic forecasts that come from different models, the same idea can easily be applied to appropriately rescaled temporal hierarchies. This relies on the fact that the probabilistic forecast at each node can be understood as coming from a different model fit to the same quantity over a time interval of constant length.

Lichtendahl, Grushka-Cockayne, and Winkler (2013) also propose an approach that averages cumulative probabilities, but find this approach to be inferior to a quantile averaging approach. Our own application of probability averaging to the reconciliation of temporal hierarchies leads to the same conclusion and these results are omitted.

3.3.3. Permuted Sample

A final alternative would be to randomly shuffle the elements within each row of $\hat{\boldsymbol{Y}}^S$. We refer to this as the 'permuted sample' $\hat{\boldsymbol{Y}}^P$. The shuffling has the effect of decoupling the dependence within each level, making the rows of $\hat{\boldsymbol{Y}}^P$ independent with respect to one another. Although this may seem to be an unreasonable approach, it provides an interesting contrast with the ranked sample and may be a useful method that guards against over-fitting when dependence is low.

3.4. Reconciliation Methods

Once the matrix $\hat{\boldsymbol{Y}}$ has been formed either as the stacked, ranked or permuted sample, it is pre-multiplied by $\boldsymbol{SP}$ to yield a reconciled sample. We consider several alternatives for $\boldsymbol{P}$ in this section. Most of these choices of $\boldsymbol{P}$ correspond to existing methods shown to have some merit in the literature on point forecast reconciliation. Their application in a probabilistic setting is novel and their relative performance is an open question that we shall investigate. The final method, which we propose in Section 3.4.4, has to the best of our knowledge never been used even for point forecast reconciliation and represents an original contribution to the literature in its own right.

3.4.1. Bottom-Up (BU)

A simple choice for $\boldsymbol{P}$ is to simply ignore information above the bottom level of the hierarchy and simply aggregate the base bottom level forecasts. For the example, in Figure 3 this implies:

$$\boldsymbol{P}_{BU} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

or more generally $\boldsymbol{P}_{BU} = \begin{bmatrix} \mathbf{0}_{m\times(M-m)} & \boldsymbol{I}_m \end{bmatrix}$, where $\mathbf{0}_{a\times b}$ denotes a $a \times b$ matrix of zeroes and $\boldsymbol{I}_a$ is an identity matrix of order a,

3.4.2. Global Average (GA)

Another method is to use information at all nodes of the hierarchy via a simple average, or

$$\tilde{\boldsymbol{Y}}_{i,\cdot} = \frac{1}{M}\sum_{j=1}^{M}\hat{\boldsymbol{Y}}_{j,\cdot} \quad \forall i.$$

This is equivalent to assuming that the matrix $\boldsymbol{P}$ is a matrix of ones scaled by $(1/M)$, that is $\boldsymbol{P}_{GA} = (1/M)\mathbf{1}_{m\times M}$. We note that the global average leads to probabilistic forecasts that are the same for every node, before being transformed back to the original scale. Simple averaging across the top and bottom levels as well as all levels of a cross-sectional hierarchy was also considered by Abouarghoub, Nomikos, and Petropoulos (2018). Simple averages outperformed the OLS combination and the bottom-up approaches in the case of energy transport forecasts.

3.4.3. Weighted Least Squares (WLS) via structural scaling

In the context of point forecasts, Athanasopoulos et al. (2017) consider reconciled forecasts as $\tilde{\boldsymbol{Y}} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}\hat{\boldsymbol{Y}}$, where $\boldsymbol{W}$ is a diagonal matrix. Although, Athanasopoulos et al. (2017) consider different weighting schemes, a simple yet effective choice of weights, particularly well suited to temporal hierarchies, is structural scaling. Here, each series is assigned a weight in line with the number of nodes aggregated to form that series. We adopt this approach here but note two distinctions between our approach and that of Athanasopoulos et al. (2017). The first distinction is a difference in notation - our definition of $\boldsymbol{S}$ already takes scaling into account. As such, using our definition of $\boldsymbol{S}$ and setting $\boldsymbol{W} = \boldsymbol{I}$ (an ordinary least squares estimator) is equivalent to a weighted least squares estimator under the definition of $\boldsymbol{S}$ used by Athanasopoulos et al. (2017). Therefore, despite setting $\boldsymbol{W} = \boldsymbol{I}$ we use the terminology 'WLS' since this is more in line with the usage of WLS in the rest of the literature. Another subtle difference between Athanasopoulos et al. (2017) and our own approach is that for the former, the element on the diagonal of $\boldsymbol{W}$ corresponding to a node in level l is set to $\boldsymbol{f}_l$ while we prefer $\boldsymbol{f}_l^2$. This reflects the fact that $\boldsymbol{W}$ is a proxy for a variance covariance matrix and that standard deviations rather than variances scale proportionally to the underlying random variable.

3.4.4. Cross-Validated (CV)

A shortcoming of many of the approaches above, including WLS with structural scaling, is that the weights are fixed. Even for a variance scaling WLS or MinT approach, weights are a function of in-sample errors and are not directly determined with reference to the objective function

ultimately used to assess forecast quality. In this section we propose a class of data-driven weights that are determined via cross-validation to maximise the sharpness of the reconciled predictive distributions, subject to calibration. The notions of sharpness and calibration are discussed by Gneiting and Katzfuss (2014). To the best of our knowledge, the use of cross-validation weights has not been considered in hierarchical reconciliation, even in the case of point forecasting.

The cross-validation procedure involves splitting the sample into three non-overlapping samples, the training sample $\mathcal{T}_{train}$, the validation sample $\mathcal{T}_{val}$ and the test sample $\mathcal{T}_{test}$. Before cross-validation, base model parameters are estimated using only training data. We denote these estimates as $\hat{\boldsymbol{\theta}}_{train}$. Then for all $t+h$ in the validation sample, a sample is produced from $\hat{F}(\boldsymbol{y}^{t+h}|\mathcal{F}^t;\hat{\boldsymbol{\theta}}_{train})$, where $\hat{F}$ is used to denote the base predictive cumulative distribution function (CDF). After pre-multiplication by some matrix $\boldsymbol{SP}$, a sample from the reconciled CDF $\tilde{F}(\boldsymbol{y}^{t+h}|\mathcal{F}^t;\hat{\boldsymbol{\theta}}_{train})$ is obtained. Let $\tilde{F}^{t+h}_{j,f_l}$ be the CDF of the margin corresponding to the j^{th} node in the level l of the hierarchy. Finally let $R(F,z)$ be a strictly proper scoring rule where F is a predictive CDF, and z is a scaled realisation.

The objective function for our cross validation sums the scores over all levels, for all series within each level and over all time points in the validation period. It is given by

$$CV(\boldsymbol{P}) = L^{-1}\sum_{l=1}^{L} CV_l(\boldsymbol{P})\,, \tag{4}$$

where

$$CV_l(\boldsymbol{P}) = (\boldsymbol{f}_1/\boldsymbol{f}_l)^{-1}\sum_{j=1}^{(\boldsymbol{f}_1/\boldsymbol{f}_l)}\sum_{t+h\in\mathcal{T}_{val}} R(\tilde{F}^{t+h}_{j,\boldsymbol{f}_l}, z^{t+h}_{j,\boldsymbol{f}_l}). \tag{5}$$

The scoring rule used for training reconciliation weights is the continuous ranked probability score (CRPS) given in general by

$$R(F,z) = \int_u (F(u) - \mathbb{1}\{z\le u\})^2 du\,, \tag{6}$$

where $\mathbb{1}\{.\}$ is an indicator function equal to 1 if the statement in braces is true and 0 otherwise. As an alternative, the log score could also be used, we prefer the CRPS since it is bounded and optimisation with respect to this score is therefore more stable. We note that the same scoring rule is used for training reconciliation weights in our empirical study but to ensure a fair comparison

between all methods a different scoring rule will be used to evaluate the final forecasts.

The quantity $CV(\boldsymbol{P})$ is optimised with respect to $\boldsymbol{P}$. Since the $\boldsymbol{P}$ matrix can be quite large we consider the following sparse structure

$$\boldsymbol{P}_{CV} = \begin{bmatrix} v_{1,1} & v_{2,1} & 0 & v_{3,1} & 0 & 0 & 0 \\ v_{1,1} & v_{2,1} & 0 & 0 & v_{3,2} & 0 & 0 \\ v_{1,1} & 0 & v_{2,2} & 0 & 0 & v_{3,3} & 0 \\ v_{1,1} & 0 & v_{2,2} & 0 & 0 & 0 & v_{3,4} \end{bmatrix}, \tag{7}$$

To define this in general, suppose we consider row i of $\boldsymbol{P}$ which corresponds to a bottom level node. Then non-zero weights are only assigned to that bottom level node and its ancestor nodes (i.e. the parent node, the parent of the parent node etc.). This structure does not use information from forecasts of sibling nodes to reconcile probabilistic forecasts. A motivation for this is that in the temporal forecasting context, the dependence within each level can be easily preserved in base forecasts. Due to the computationally intensive nature of cross-validation, for the case study in Sections 4 and 5, we consider a variation of the structure above that optimises over fewer weights. In particular the same value is used for all weights corresponding to the same level, giving the following $\boldsymbol{P}$ matrix for the hierarchy in Figure 3:

$$\boldsymbol{P}_{CVR} = \begin{bmatrix} v_1 & v_2 & 0 & v_3 & 0 & 0 & 0 \\ v_1 & v_2 & 0 & 0 & v_3 & 0 & 0 \\ v_1 & 0 & v_2 & 0 & 0 & v_3 & 0 \\ v_1 & 0 & v_2 & 0 & 0 & 0 & v_3 \end{bmatrix}, \tag{8}$$

where v_l corresponds to the weight on all the nodes in the level l. Note that the bottom-up method in Section 3.4.1 is a special case of this structure, where $v_L = 1$ and all other weights are zero.

We consider three cases for the weights: (1) all weights in a row sum to one and are positive; (2) all weights in a row sum to one; and (3) all weights are unconstrained. Ensuring all weights sum to one guarantees that unbiased base forecasts will also be unbiased after reconciliation and is a property shared by the reconciliation matrix for bottom-up and WLS. Ensuring that all weights are positive as well potentially leads to a more stable algorithm (highly correlated series cannot be assigned weights that cancel one another out). Perhaps more crucially, restricting all weights to be

positive ensures the sample of points from the reconciled density forecasts are non-negative as long as the sample of points from the base density forecasts are non-negative.

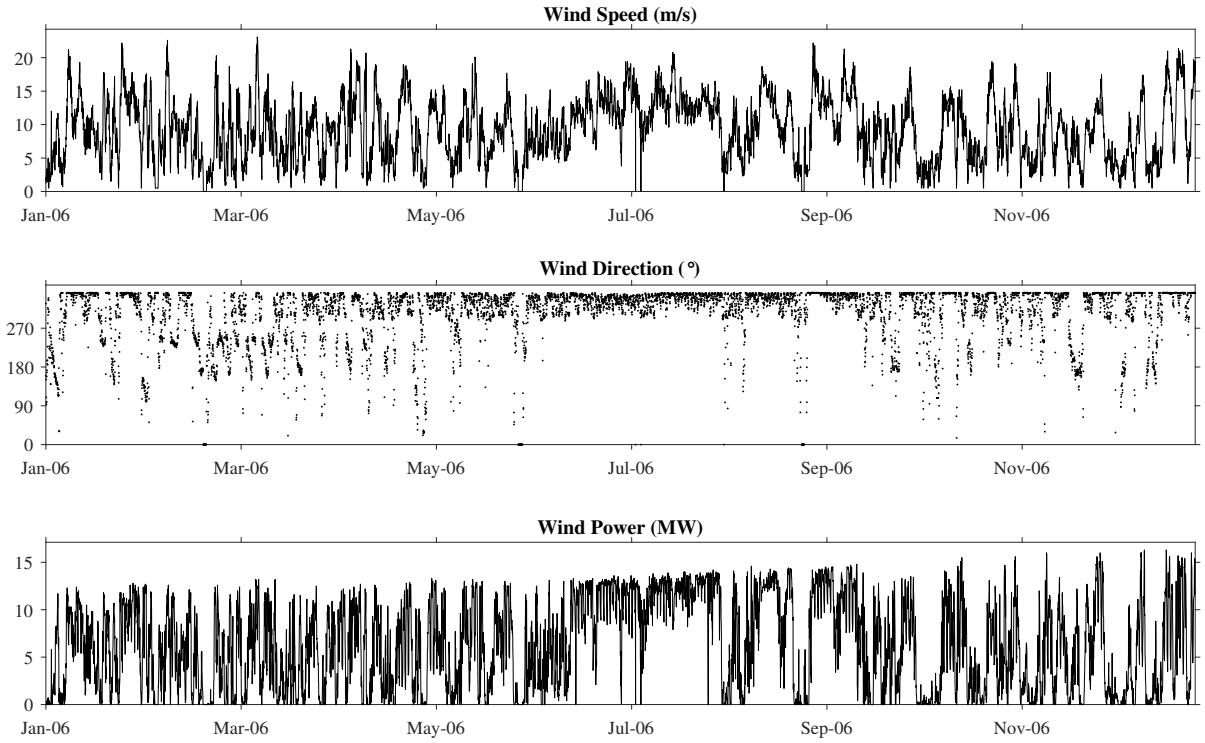

Figure 5: Hourly time series of wind speed, wind direction and wind power in the Rokas wind farm, Crete.

4. Empirical design

As a case study of the methods we propose in Section 3, we consider three datasets, the first two are wind power data, while the third is data on electric load. In all three cases, spot power exchange markets are typically a day-ahead auction, but the market price is calculated for each hour of the following day. The nature of this market thus lends itself to a hierarchical approach using a daily cycle and an hourly frequency at the bottom level. As such we construct a temporal hierarchy with $\boldsymbol{f} = (24, 12, 8, 6, 4, 3, 2, 1)'$ and the following $\boldsymbol{S}$ matrix

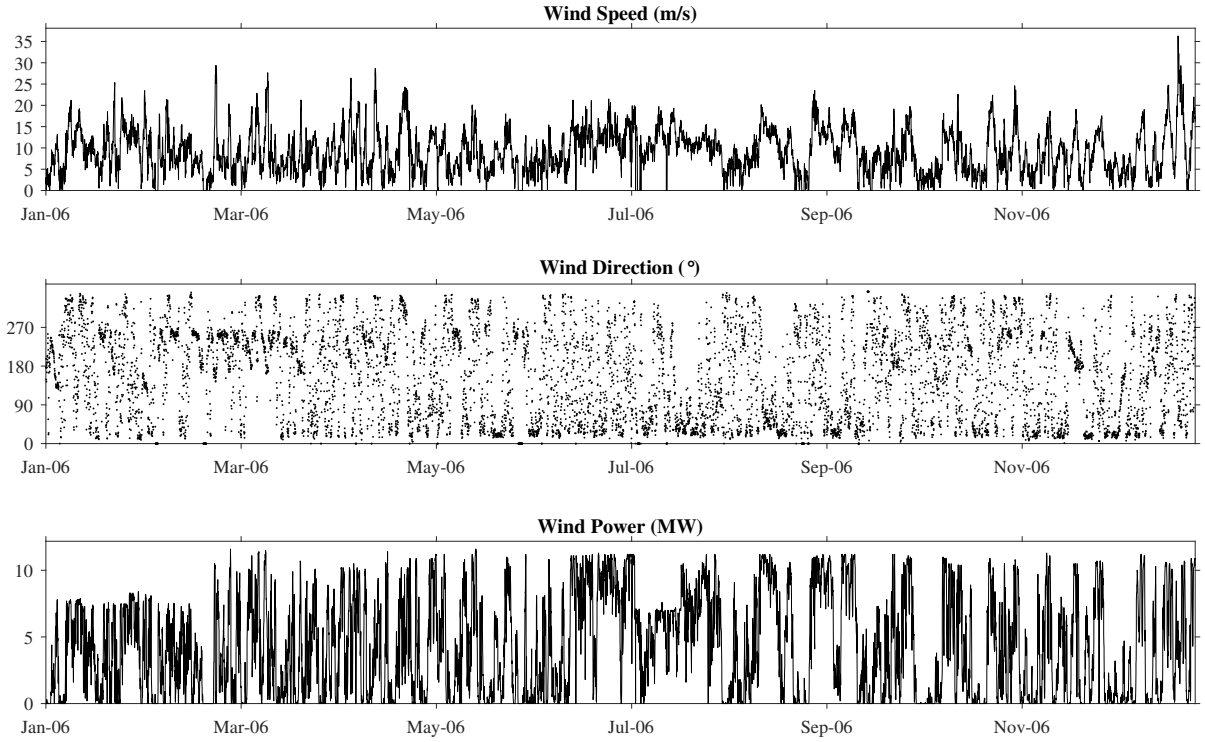

Figure 6: Hourly time series of wind speed, wind direction and wind power in the Aeolos wind farm, Crete.

$$\boldsymbol{S} = \begin{bmatrix} 24^{-1}\boldsymbol{\iota}'_{24} \\ 12^{-1}\boldsymbol{I}_2 \otimes \boldsymbol{\iota}'_{12} \\ 8^{-1}\boldsymbol{I}_3 \otimes \boldsymbol{\iota}'_{8} \\ 6^{-1}\boldsymbol{I}_4 \otimes \boldsymbol{\iota}'_{6} \\ 4^{-1}\boldsymbol{I}_6 \otimes \boldsymbol{\iota}'_{4} \\ 3^{-1}\boldsymbol{I}_8 \otimes \boldsymbol{\iota}'_{3} \\ 2^{-1}\boldsymbol{I}_{12} \otimes \boldsymbol{\iota}'_{2} \\ \boldsymbol{I}_{24} \end{bmatrix}, \tag{9}$$

where $\boldsymbol{\iota}_a$ is an column of a ones and $\otimes$ denotes the Kronecker product. The overlapping hierarchy consists of 1×24 hour forecast, 2×12 hourly forecasts, 3×8 hourly forecasts, 4×6 hourly forecasts, 6×4 hourly forecasts, 8×3 hourly forecasts, 12×2 hourly forecasts and 24×1 hourly forecasts. This amounts to $L = 8$ levels, $M = 60$ nodes and $m = 24$ bottom-level nodes in the hierarchy. Forecasts of each series can be of interest due to different operational purposes. For

instance, Fan and Hyndman (2011), Pinson (2013) and Hong and Fan (2016) explain that forecasts up to 2 hours ahead are crucial for dispatch, real time generation control, reliability analysis, spinning reserve allocation and load shifting by electricity retailers, while much longer lead times up to several days ahead are also relevant to decision-making for unit commitment, transmission operations, load-balancing and scheduling for spinning reserve, maintenance planning for generators and planning for optimal trading strategies. Also, although we focus on probabilistic forecasts of wind power and electric load up to 24 hours ahead, the method we proposed can be still applied to multi-day ahead forecasts.

4.1. Wind Power Data

It is a major challenge for grid operators to maximise the utilisation of wind power due to the intermittent nature of the generation. Due to the inherent uncertainty in the wind power forecasting, probabilistic approaches have received increasing attention recently (Taylor, 2017; Roulston and Smith, 2003; Gneiting et al., 2006; Jeon and Taylor, 2012; Hering and Genton, 2010; Taylor and Jeon, 2015; Dowell and Pinson, 2015). Probabilistic forecasts enable more informed decision-making by allowing for the optimal design of bidding strategies and power balance by wind farm operators, grid system operators and electricity traders (Pinson, 2013). The wind power data we consider come from the Rokas and Aeolos wind farms in Crete, the largest island in the Aegean Sea. The island has an autonomous electricity grid and high wind energy potential. The generation capacities of the Rokas and Aeolos wind farms were 16.3MW and 11.6MW, respectively, in 2006.

Due to the highly non-linear evolution of wind power, rather than model and forecast this series directly, we follow an indirect approach shown to be more accurate by Jeon and Taylor (2012). Under this approach, data are collected on exogenous drivers of wind power, for example wind speed. Forecasts of wind speed are produced and converted to wind power via a 'power curve' function. Jeon and Taylor (2012) do not use the deterministic power curve provided by turbine manufacturers since there is some residual noise that can be attributed to other factors. These include: changes in wind direction, air pressure, temperature, precipitation, the complexity of the terrain, different behaviour between speed up and down, turbulence in the turbines, maintenance of turbines, and errors in measurement. To account for the stochastic nature of the conversion from wind speed to wind power, Jeon and Taylor (2012) instead propose using a kernel density estimate

of wind power conditional on wind speed. This approach still requires the selection of a univariate model for forecasting wind speed.

An advantage of the approach of Jeon and Taylor (2012) is that it can easily be extended to consider an additional exogenous driver, namely wind direction. This requires the selection of a bivariate model for forecasting wind speed and wind direction that are ultimately converted to wind power forecasts. Both univariate (wind speed only) and bivariate (wind speed and direction) candidate models are described in Section 4.3. Wind speed and direction recorded at the turbine hub height of the two wind farms as well as wind power are plotted in Figures 5 and 6. Data are recorded for each hour in 2006, which amounts to 8,760 observations in each series. Figures 5 and 6 show that wind power is more volatile than wind speed, and the volatilities tend to be clustered. Figure 7 illustrates the 24 hourly, 8 hourly and 2 hourly time series of Rokas, aggregated from the 1 hourly time series. A lower frequency time series exhibits more smoothed movements. We observe similar patterns for Aelos.

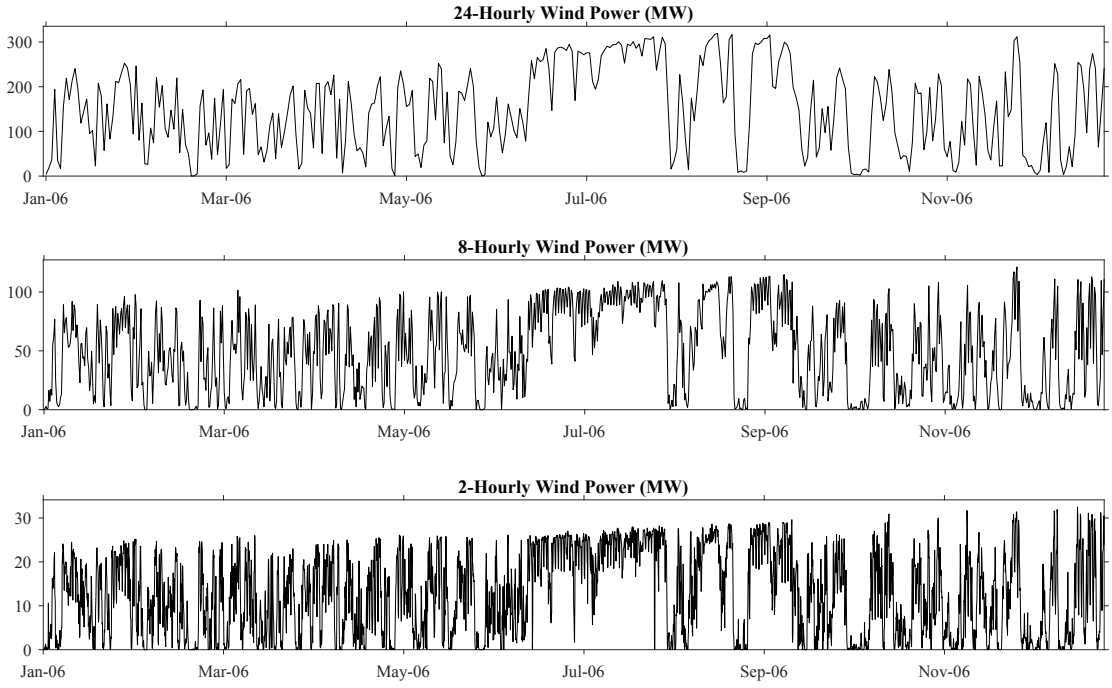

Figure 7: The 24 hourly, 8 hourly and 2 hourly time series of wind power in the Rokas wind farm, Crete.

Each time series was split to $\mathcal{T}_{train}$, a 6-month training period from 1 January 2006 to 30 June 2006, used to train models for wind speed and direction; $\mathcal{T}_{val}$, a 3-month validation from 1 July 2006 to 30 September 2006, used to choose the most accurate model for each level in the temporal

hierarchy and to select the cross-validation weights in Section 3.4.4; and $\mathcal{T}_{test}$, a 3-month test period from 1 October 2006 to 31 December 2006, reserved for evaluation of reconciliation methods.

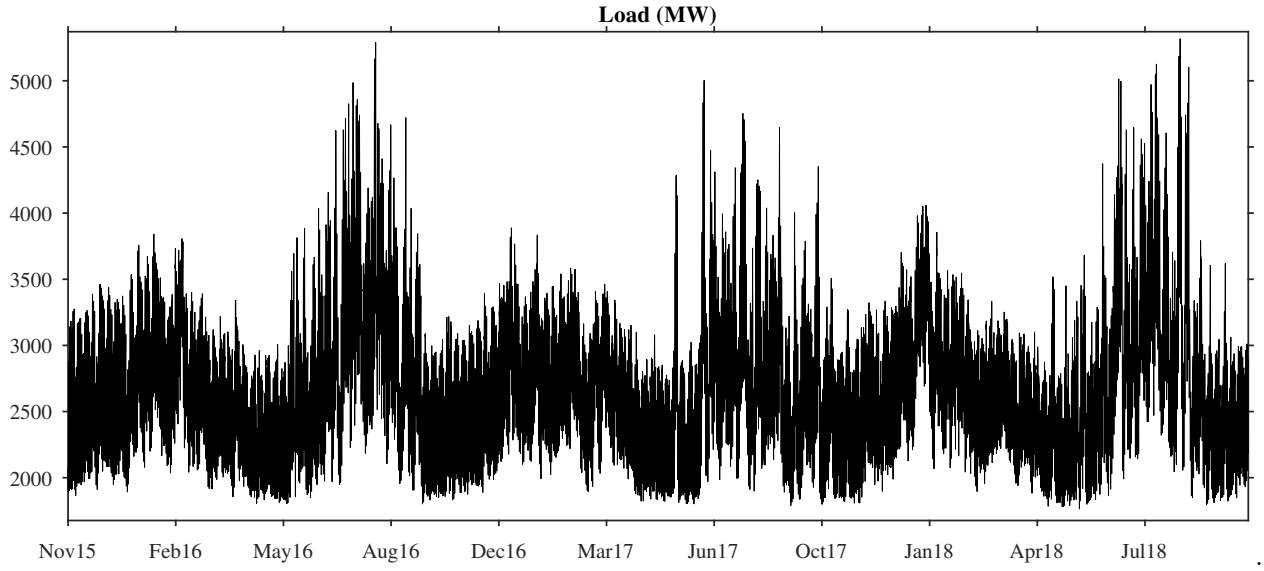

Figure 8: Hourly time series of electric load in Boston, US.

4.2. Electric Load Data

Another case study used in the empirical study relates to electric load. For a discussion on the importance of probabilistic forecasts for electric load forecasting see Hong and Fan (2016), Hyndman and Fan (2010), Ben Taieb, Taylor, and Hyndman (2017) and references therein. We note that Nystrup et al. (2018) has applied temporal forecast reconciliation methods to electric load data but only in the point forecasting setting. The following empirical example therefore represents the first attempt to bring these two strands of the literature together. Our data is the hourly time series of electric load for Boston, MA (load zone code: NEMA), downloaded from ISO New England. The time series is one of the data sets used in GEFCom2017 (Hong, Xie, and Black, In press). The period used in our empirical analysis ranges from 1 November 2015 to 31 October 2018, returning 26,304 observations. We divided the data into three periods: $\mathcal{T}_{train}$ from 1 November 2015 to 31 October 2016; $\mathcal{T}_{val}$ from 1 November 2016 to 31 October 2017; and $\mathcal{T}_{test}$ from 1 November 2017 to 31 October 2018. To account for special day effects, the public holidays in this period, published by the US Office of Personnel Management, were encoded as an exogenous dummy variable and used in the base forecasting models.

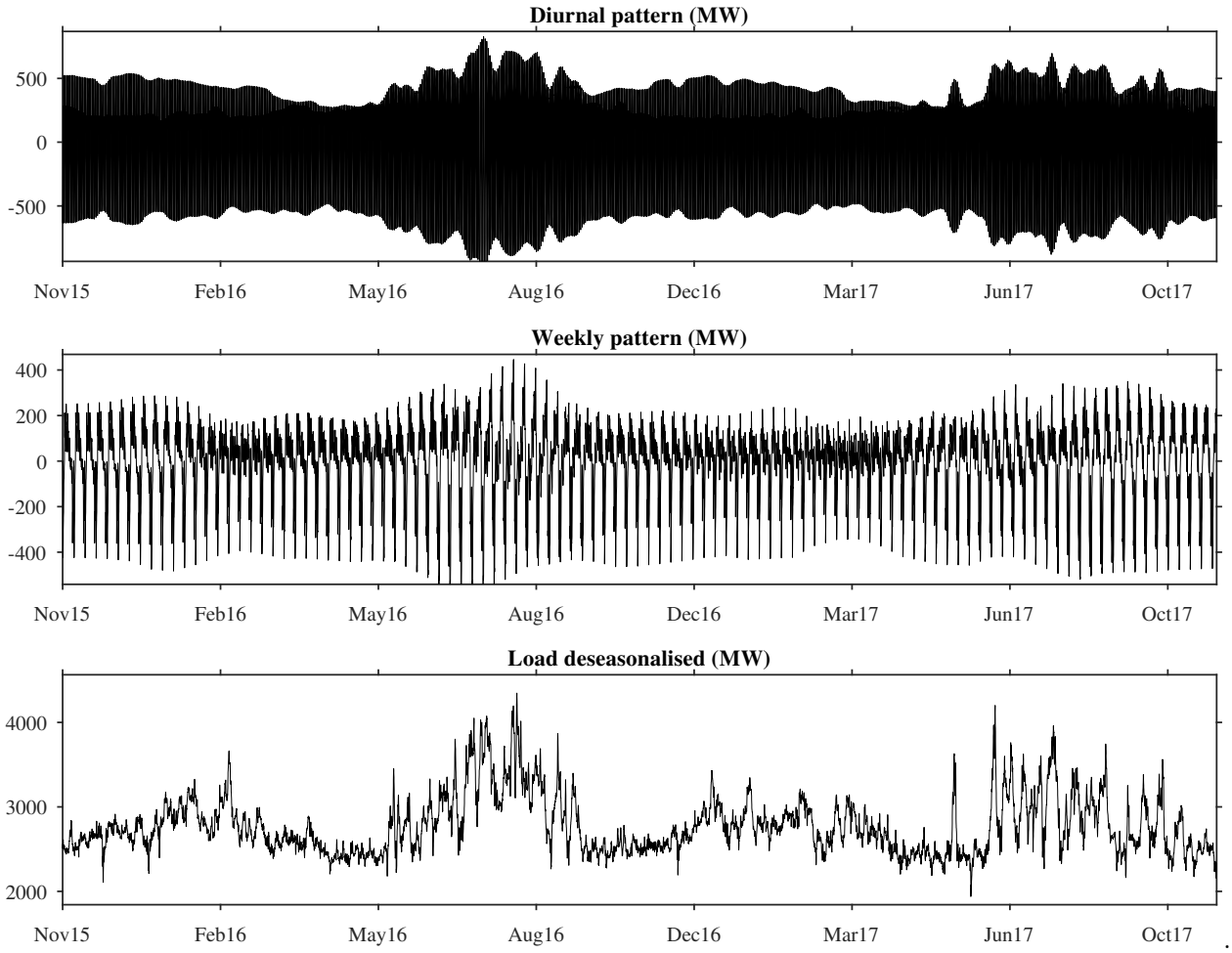

Figure 9: Hourly time series of diurnal and weekly patterns in the in-sample period, identified by the mstl function in the forecast package in R, and the corresponding deasonalised load series.

As shown in Figure 8, the load time series has strong yearly, weekly and diurnal patterns (Hyndman and Fan, 2010; Roach, In press). The peak loads are not expected to be captured accurately in a multiple temporal aggregation framework, such as temporal hierarchies, due to the shrinkage on the seasonal component (Spiliotis et al., 2018). For this reason, instead of fitting our base density forecasting models directly to the original time series, we fit the models to the seasonally-adjusted time series. The weekly and diurnal patterns were captured on a rolling basis in the data before each forecast origin in $\mathcal{T}_{val}$ and $\mathcal{T}_{test}$, using the mstl function in the forecast package in R (Hyndman, 2014), which is discussed in detail in Section 6.8 of Hyndman and Athanasopoulos (2018). This method decomposes a series into additive seasonal components. Letting x_t be hourly

load at time t, $x_t = d_t + w_t + e_t$, where d_t and w_t are the daily and weekly seasonal components respectively and e_t is a residual component. We fit our base model to e_t using the information only available in the in-sample period. In the evaluation period, the two seasonal components are updated on a rolling basis using information up to the forecast origin, $t-1$. The forecasts for d_t and w_t are obtained from a seasonal naive forecast with a 52-week yearly cycle so that we take into account daily, weekly and yearly seasonality. Once we obtain density forecasts for e_t, we reconstruct the predictive density in the original scale by adding back the forecasts for the daily and weekly seasonality. The additive nature of this method preserves the coherence of the density forecasts. We found this procedure returns markedly better forecasting results in our empirical study compared to working directly with the original time series. The top two panels of Figure 9 show d_t and w_t respectively while the bottom panel shows the deseasonalised data. Alternative methods for deaseasonalising the data could have been used, with the important caveat that the deaseasonalised data must be coherent in precisely the same way as the raw data. This is a critical precondition for using reconciliation techniques, and we do not recommend using reconciliation if data is deseasonalised in a way that does not preserve coherence (for example multiplicative seasonality).

4.3. Base Probabilistic Forecasting Models for Wind Power and Electric Load

Table 1: Models chosen for each wind farm and each hierarchical level. Univariate models produce wind speed density forecasts only and convert these to power forecasts. Bivariate models produce density forecasts of wind speed and wind direction before converting these to power forecasts.

Level	**Base forecast model for Rokas**	**Base forecast model for Aeolos**
24 hourly	Univariate ARFIMA-FIGARCH with Gaussian	Univariate ARMA-FIGARCH with Student t
12 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
8 hourly	Bivariate VARMA-GARCH with Student t	Univariate ARMA-FIGARCH with Student t
6 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
4 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
3 hourly	Bivariate VARMA-GARCH with Student t	Bivariate VARMA-GARCH with Student t
2 hourly	Univariate ARMA-GARCH-skew t	Bivariate VARMA-GARCH with Student t
1 hourly	Univariate ARMA-FIGARCH with skew t	Univariate ARMA-GARCH with skew t

Before addressing the issue of of probabilistic forecast reconciliation it is first necessary to determine base forecasting models. Recall that for wind power, we use an indirect approach where we forecast wind speed and possibly wind direction before converting these to power forecasts. For wind speed and direction, statistical models are considerably cheaper than a numerical weather pre-

Table 2: Models chosen for each hierarchical level for electric load in Boston.

Level	**Base forecast model for Boston**
24 hourly	ARFIMA-GARCH with Student t
12 hourly	ARFIMA-FIGARCH with Gaussian
8 hourly	ARMA-FIGARCH with Student t and Box-Cox transformation
6 hourly	ARMA-GARCH with Student t and Box-Cox transformation
4 hourly	ARMA-GARCH with Gaussian and Box-Cox transformation
3 hourly	ARMA-FIGARCH with Student t and Box-Cox transformation
2 hourly	ARMA-GARCH with Student t
1 hourly	ARMA-FIGARCH with Gaussian

diction (NWP) system (Sloughter, Gneiting, and Raftery, 2010) while being very competitive for short lead times (Pinson, 2013). The models we consider include univariate autoregressive moving average − generalized autoregressive conditional heteroskedasticity (ARMA−GARCH) models for forecasting wind speed alone as well as their VEC−type bivariate equivalent (VARMA−GARCH) models (Bollerslev, Engle, and Wooldridge, 1988) for forecasting both wind speed and wind direction. We also model the long memory dependence in the mean and the volatility of the wind speed time series using the autoregressive fractionally integrated moving average model (ARFIMA; Granger and Joyeux, 1980; Hosking, 1981) and the fractionally integrated generalized autoregressive conditionally heteroskedastic model (FIGARCH; Baillie, Bollerslev, and Mikkelsen, 1996) respectively. For deseasonalised electric load, we also consider the univariate statistical models described above. Univariate models are commonly used in the literature when forecasting short-term energy load (Taylor, 2003; Nystrup et al., 2018).

We fit all contender models to $\mathcal{T}_{train}$ with Gaussian, Student t and skew t distribution assumptions for the noise term. To account for a potentially high degree of non-normality, we also considered a Box-Cox transformation (Box and Cox, 1964). In all cases, 1,000 Monte-Carlo simulated sample paths were generated to construct 1 to 24 hour ahead density forecasts for each level of the hierarchy from each forecast origin in $\mathcal{T}_{test}$. We opted not to re-estimate model parameters as we rolled the forecast origin forward, as it was impractical due to high computational cost. All models were evaluated over $\mathcal{T}_{val}$ using the average value of CRPS across all horizons (1 to 24 hours ahead).

If we look at the models selected for wind power first, the best model for each hierarchical level is presented in Table 1. It is notable that Student t and skew t were selected more frequently compared to the Gaussian distribution, which indicates that the conditional distribution of wind speed is non-

Gaussian. In line with Taylor, McSharry, and Buizza (2009), the univariate model of wind speed with fractional integration in level and volatility was found to produce the most accurate density forecasts for daily (24 hourly) wind speed forecasts. For other frequencies, the bivariate VARMA-GARCH model was frequently identified as the best for both wind farms, indicating for these hierarchical levels, using wind direction as a driver of wind power produces more accurate density forecasts. Regarding the base models chosen for load forecasting, the most accurate density forecast models chosen in $\mathcal{T}_{val}$ for each temporal hierarchical level was shown in Table 2. Skewness and excess kurtosis of the load time series were higher than for the wind time series. As a consequence models with a Box-Cox transformation were chosen more often for the load data.

All computation was carried out on a standard desktop computer with an Intel i7-7700 CPU and without parallel processing. The following computation times correspond to the case of the electric load of Boston, which is the longest time series among our data sets. Overall, it takes 16 minutes to fit the models chosen in Tables 1 and 2 at all 8 frequencies and a further 9.2 seconds to produce base density forecasts at all levels. Training the reconciliation weights via cross validation is much slower taking up to 6.2 hours (for the case where the permuted sample is used with the constraint that all reconciliation weights are positive and sum to one). Once these weights have been determined, or for methods based on known weights (such as WLS), reconciliation of base forecasts is almost instantaneous, taking only 0.04 seconds. Taking all of this into account, in practice, when forecasts are required at a very short lead time, estimated parameters of the base forecasting models and the reconciliation weights should be calculated once then retained. We have followed this approach in implementing our rolling window evaluation.

5. Empirical results

5.1. Cross-validation Weights

The weights for the cross-validated method described in Section 3.4.4 are calculated based on forecast evaluation over $\mathcal{T}_{val}$. The weights for each of the three different constraints using the $\boldsymbol{P}_{CVR}$ matrix are presented in Table 3 for the Rokas wind farm, Table 4 for the Aeolos wind farm and Table 5 for the electric load of Boston. It is noteworthy that there is a tendency for high weights (in absolute value) on the 1 or 2 hourly hierarchical level for Rokas and Aeolos, while no such pattern is observed for the load data. This is sensible as wind power data are characterised by a high degree

of intermittence with potentially strong signals at higher frequencies. Within the framework we consider, and with the caveat that different sampling methods and constraints on the weights lead to differences in forecast accuracy, there is some evidence that high-frequency data is informative for all levels of the hierarchy for the wind power data.

Table 3: Weights(v) of the CV method in Section 3.4.4 derived for Rokas, determined by minimising the average of the level-wise average CRPS values in the hierarchy. The sum of v is the row sum.

Sampling scheme	**Hierarchical level**								
Method	24h	12h	8h	6h	4h	3h	2h	1h	Sum
Permuted Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00
$\sum v_i = 1$	-0.37	0.05	0.38	-0.15	0.19	0.10	0.87	-0.07	1.00
Unconstrained	-0.28	-0.03	0.44	-0.19	0.20	0.09	0.87	-0.05	1.05
Ranked Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.00	0.01	0.00	0.02	0.00	0.98	0.00	1.00
$\sum v_i = 1$	-0.34	0.29	0.23	-0.07	0.49	-0.38	0.64	0.14	1.00
Unconstrained	-0.25	-0.08	0.62	-0.11	0.53	-0.61	0.94	-0.04	1.00
Stacked Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.78	1.00
$\sum v_i = 1$	-0.01	-0.04	0.04	-0.02	0.03	0.07	0.38	0.56	1.00
Unconstrained	-0.01	0.00	0.05	-0.03	0.08	0.24	0.01	0.35	0.69

Table 4: Weights(v) of the CV method in Section 3.4.4 derived for Aeolos, determined by minimising the average of the level-wise average CRPS values in the hierarchy. The sum of v is the row sum.

Sampling scheme	**Hierarchical level**								
Method	24h	12h	8h	6h	4h	3h	2h	1h	Sum
Permuted Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.25	1.00
$\sum v_i = 1$	-0.01	-0.00	0.53	-0.34	-0.00	0.34	-0.19	0.68	1.00
Unconstrained	0.06	-0.23	0.46	-0.43	-0.09	0.18	0.00	0.86	0.82
Ranked Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.84	1.00
$\sum v_i = 1$	0.17	0.14	0.16	-0.34	-0.40	0.48	-0.03	0.82	1.00
Unconstrained	0.18	-0.13	0.24	-0.34	-0.59	0.72	-0.00	0.78	0.86
Stacked Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.99	1.00
$\sum v_i = 1$	0.01	-0.01	0.03	-0.05	-0.01	0.02	0.02	1.00	1.00
Unconstrained	-0.01	-0.02	0.03	-0.07	0.13	0.04	0.32	0.34	0.77

Table 5: Weights(v) of the CV method in Section 3.4.4 derived for Boston, determined by minimising the average of the level-wise average CRPS values in the hierarchy. The sum of v is the row sum.

Sampling scheme	**Hierarchical level**								
Method	24h	12h	8h	6h	4h	3h	2h	1h	Sum
Permuted Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.00	0.82	0.00	0.00	0.00	0.18	0.00	1.00
$\sum v_i = 1$	-0.20	0.18	0.60	-0.31	0.47	0.17	0.32	-0.22	1.00
Unconstrained	0.26	0.16	-0.10	-0.42	0.68	0.34	0.46	-0.38	1.00
Ranked Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.07	0.12	0.06	0.17	0.20	0.11	0.15	0.13	1.00
$\sum v_i = 1$	0.07	0.12	0.06	0.17	0.20	0.11	0.15	0.13	1.00
Unconstrained	0.01	0.16	0.09	0.13	0.22	0.14	0.12	0.14	1.01
Stacked Sample									
$\sum v_i = 1$ & $\forall v_i \geq 0$	0.00	0.08	0.00	0.00	0.89	0.00	0.03	0.00	1.00
$\sum v_i = 1$	0.01	0.23	-0.04	0.06	0.75	-0.34	-0.03	0.36	1.00
Unconstrained	-0.07	0.22	0.57	-0.58	0.30	0.33	0.06	0.16	1.00

5.2. Forecast evaluation

We now turn our attention to evaluating the forecast performance of all the methods suggested in Section 3 over $\mathcal{T}_{test}$, for all three datasets. To evaluate the accuracy probabilistic forecasts of the hierarchy as a whole, we used the energy score, a multivariate scoring rule introduced by Gneiting and Raftery (2007) that generalises the CRPS. This score is used to evaluate all density forecasts, including base forecasts, reconciled benchmarks and the CV method. The use of energy score should be distinguished from the score used in Equation 5 with the latter used to train the cross validation weights but not evaluate forecasts. Let $(\breve{\boldsymbol{y}}_1^t, \ldots, \breve{\boldsymbol{y}}_N^t)'$ and $(\breve{\boldsymbol{y}}_1^{t*}, \ldots, \breve{\boldsymbol{y}}_N^{t*})'$ be samples of vectors independently drawn from a multivariate predictive density $\breve{F}(\boldsymbol{y})$ and $\boldsymbol{y}^t$ be the eventual realisation of the vector that is the target of the forecast. The breve notation is used to denote that the sample from the predictive density can be either base forecasts ($\hat{\boldsymbol{y}}$) or reconciled forecasts ($\tilde{\boldsymbol{y}}$). The energy score is given by

$$ES(\breve{F}(\boldsymbol{y}^t), \boldsymbol{y}^t) = \sum_{i=1}^{N} ||\breve{\boldsymbol{y}}_i^t - \boldsymbol{y}^t|| - \frac{1}{2}\sum_{i=1}^{N} ||\breve{\boldsymbol{y}}_i^t - \breve{\boldsymbol{y}}_i^{t*}||\,, \tag{10}$$

where $||\boldsymbol{v}|| = \sqrt{\sum v_i^2}$ denotes the Euclidean norm. Energy score was chosen since, as a multivariate score it takes into account both the margins and dependence structure of the multivariate density forecast. In addition to computing the energy score for the entire hierarchy, we also consider looking

at the energy score for the bottom level series ($\boldsymbol{z}_8$) and the top level series, ($\boldsymbol{z}_1$). In the latter case, energy score is equivalent to CRPS.

For all possbile combinations of reconciliation methods (defined in Section 3.4) and sampling scheme (defined in Section 3.3), the energy scores of the density forecasts of each hierarchy (60 nodes) are evaluated for all $t \in \mathcal{T}_{test}$. These values are averaged across $\mathcal{T}_{test}$, to be presented in Table 6 for each data set. Lower values of this score indicate more accurate forecasts.

Table 6: Energy score of all the nodes in each hierarchy

Data:	**Rokas**			**Aeolos**			**Boston**		
Sampling Scheme:	Permuted	Ranked	Stacked	Permuted	Ranked	Stacked	Permuted	Ranked	Stacked
Base	86.5	84.2	86.5	72.3	69.6	72.2	6437	6638	6420
Bottom-up	79.9	74.0	79.7	67.8	60.1	67.7	6944	5951	5963
Global Average	93.5	72.1	93.4	79.8	60.0	79.9	4571	3889	4097
WLS	85.4	72.7	85.4	72.6	60.1	72.6	4200	3881	3958
Cross-validated									
- $\sum v_i = 1$ & $\forall v_i \geq 0$	75.6	71.5	83.3	62.7	59.8	67.8	3910	3993	**3796**
- $\sum v_i = 1$	**70.7**	71.6	82.1	61.5	59.6	67.3	4375	3993	4843
- Unconstrained	71.3	72.0	100.3	60.3	**59.0**	74.3	5197	4214	4390

Note: Lower values are better. The best value in each of the Rokas and Aeolos wind farms and the electric load of Boston is in bold.

We also tested the statistical significance of the differences in performance for every combination of approaches in terms of multiple comparisons from the best (MCB), which focuses on the average (across origins) ranks. In Figure 10 we plot the mean ranks of each approach and sampling scheme together with their intervals. The intervals are calculated based on studentised range and are formed arbitrarily as the mean rank plus/minus half of the studentised range (for more details, see Koning et al., 2005). The results of each data set are presented in different panels. The best method for each data set is depicted at the bottom of the respective panel, with the worst-performing method presented at the top of each panel. To distinguish between the three sampling schemes, the permuted, ranked and stacked cases are respectively depicted as red, blue an green. The difference between the forecast performance of two methods is statistically significant if their intervals do not overlap. The grey-shaded area denotes the range covered by the intervals of the best method in each case, so if the intervals of another method fall in this area, then that method is not significantly different than the best one.

Based on the results from Table 6 and Figure 10 we make a number of conclusions. From Table 6 we observe that the base forecasts (i.e. no reconciliation) are usually amongst the poorest performing methods. Indeed, Figure 10 shows that the gains in performance from using the best

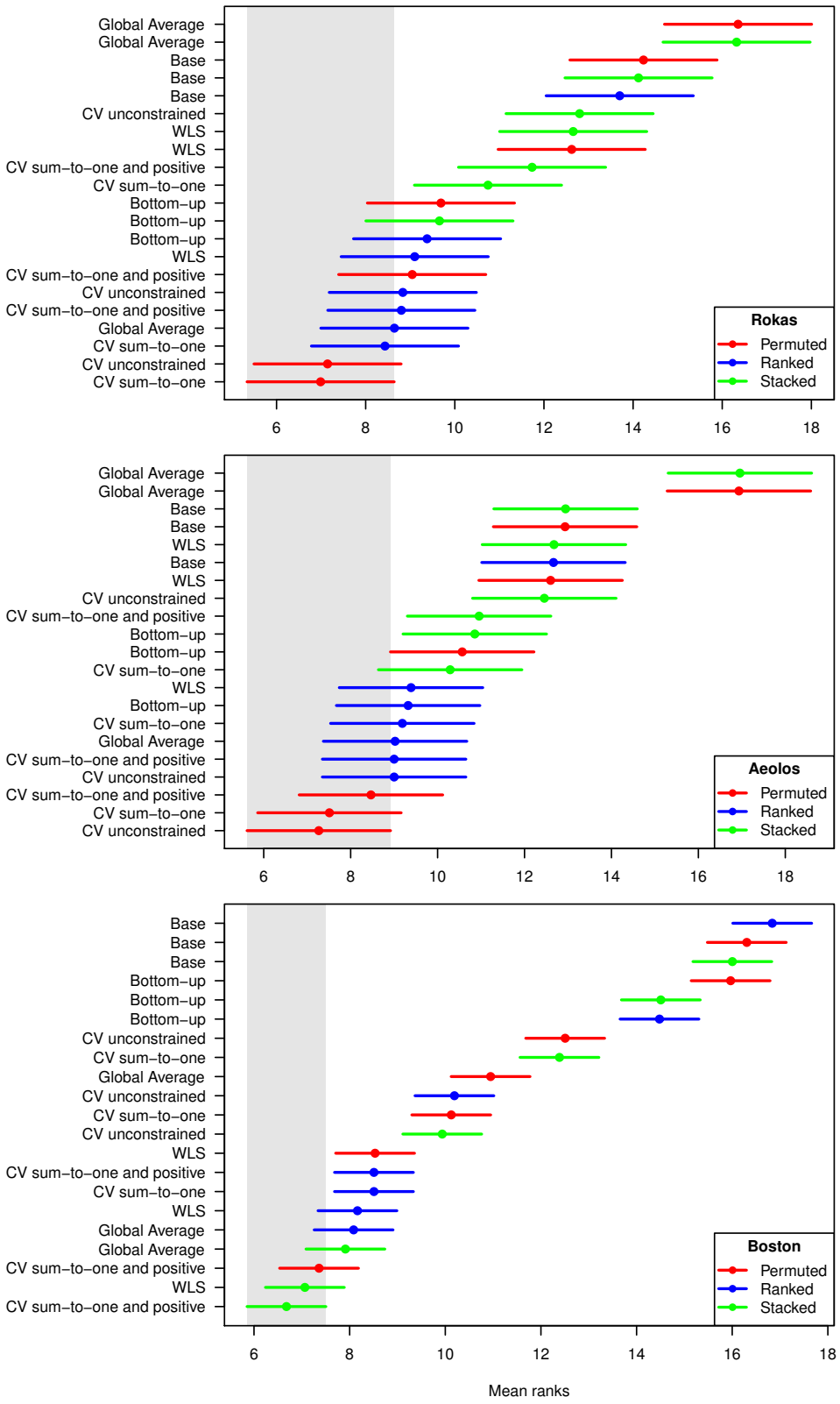

Figure 10: Multiple comparisons from the best. Intervals are given by the mean rank plus/minus half of the studentised range. Methods with a range lying entirely outside the grey bands have a significantly poorer forecasting performance than the best method.

approaches in each case are statistically significant compared to base. Furthermore, all reconciliation approaches are significantly better than bottom-up (as well as base) for energy load, irrespective of the sampling scheme. These results establish that forecast reconciliation methods that combine information from base forecasts at all levels can improve forecast accuracy in the probabilistic setting. This novel result generalises similar well-established results for point forecasting.

Table 6 also highlights the best performing combination of method and sampling scheme for each dataset in bold. In all cases these are methods that determine the reconciliation weights using cross validation rather than global average or WLS. Moreover, the performance of global average and WLS (permuted and stacked sampling schemes) are also significantly worse than the best approaches for wind energy. This novel result suggests that the optimality properties established theoretically for point forecast reconciliation do not carry over to the probabilistic case.

Finally, the importance of correctly selecting a sampling scheme to construct joint base forecasts prior to reconciliation is apparent. For wind energy output, permuted and ranked sampling schemes generally dominate the stacked, while the opposite is true for energy load. As long as reconciliation is used, the ranked sample provides the best result overall in the sense that for each dataset nearly all methods using a ranked sample cannot be statistically distinguished from the best method.

5.3. Level-specific results

While the results discussed so far evaluate the methods on the basis of the entire temporal hierarchy, in some cases only forecasts at a specific frequency are of interest. In light of this we break down the results to focus on individual levels in isolation. Table 7 provides the energy score (identical to CRPS) for the forecast wind power/electric load over the entire 24 hourly period. Table 8 provides the energy score for the forecast wind power/electric load for the next 24 1-hourly forecasts (since only the bottom level is evaluated - bottom-up results are identical to base). The general results are similar to those seen for looking at the overall hierarchy. Reconciliation improves forecast accuracy relative to base and bottom-up forecasts; there is merit in taking a CV approach relative to WLS; and the ranked sample offers the most robust sample scheme when used in conjunction with reconciliation.

In Figure 11 we plot the CRPS values of base forecasts, WLS using ranked sample and cross-validated using ranked sample with sum-to-one weights over different horizons. Although the three months in the evaluation period is not sufficient to obtain smooth lines of CRPS in the plots, there

Table 7: Energy score of the top node: 24 hourly

Data	**Rokas**			**Aeolos**			**Boston**		
Sampling Scheme	Permuted	Ranked	Stacked	Permuted	Ranked	Stacked	Permuted	Ranked	Stacked
Base	45.5	45.5	45.4	36.2	36.0	36.1	2512	2518	2512
Bottom-up	41.8	35.9	41.7	34.9	28.8	34.9	4158	3219	3285
Global Average	44.5	34.3	44.4	38.9	29.2	39.0	2290	1977	2021
WLS	44.5	34.3	44.4	38.9	29.2	39.0	2290	1977	2021
Cross-validated									
- $\sum v_i = 1$ & $\forall v_i \geq 0$	37.3	32.8	43.3	31.0	28.5	34.9	1983	2030	**1877**
- $\sum v_i = 1$	**32.5**	33.3	41.4	30.1	28.5	34.5	2150	2030	2525
- Unconstrained	33.5	34.2	55.5	29.1	**27.7**	39.9	2768	2159	2208

Note: Lower values are better. The best value in each of the Rokas and Aeolos wind farms and the electric load of Boston is in bold.

Table 8: Energy score of the bottom nodes: 1 hourly

Data	**Rokas**			**Aeolos**			**Boston**		
Sampling Scheme	Permuted	Ranked	Stacked	Permuted	Ranked	Stacked	Permuted	Ranked	Stacked
Base	12.1	12.2	12.1	10.0	9.8	10.0	919	895	880
Bottom-up	12.1	12.2	12.1	10.0	9.8	10.0	919	895	880
Global Average	15.3	12.0	15.3	12.7	9.8	12.7	708	597	641
WLS	12.2	12.1	12.2	10.1	**9.7**	10.1	614	604	604
Cross-validated									
- $\sum v_i = 1$ & $\forall v_i \geq 0$	11.9	12.0	12.2	9.9	9.8	10.0	599	610	**594**
- $\sum v_i = 1$	**11.7**	12.0	12.1	**9.7**	**9.7**	9.9	678	610	729
- Unconstrained	**11.7**	12.0	14.3	**9.7**	**9.7**	10.6	781	638	666

Note: Lower values are better. The best value in each of the Rokas and Aeolos wind farms and the electric load of Boston is in bold.

is a clear tendency for the CRPS values to increase with forecast lead times in each plot. The title of each plot in Figure 11 indicates the average improvement of cross-validated over base, in terms of CRPS, where lower values are preferred. For example, the 24 hourly density forecast of cross-validated produced CRPS values that are 27.0%, 21.0% and 19.0% smaller than base in Rokas, Aeolos and Boston, respectively. For the two wind farms, as we increase the forecast resolution by moving further down the plot, this enhancement tended to be reduced. This indicates that wind power density forecasts at the higher resolution are enhanced by synthesizing forecasts at lower resolutions. For electric load, the highest gain was for 12 hourly, followed by 1 hourly. Overall, these results demonstrate how reconciliation is able to 'hedge' against misspecification error at all levels by synthesizing information across all hierarchical nodes.

6. Concluding Comments

This paper introduced methodology for the reconciliation of probabilistic forecasts. Despite a particular focus on temporal hierarchies, the method can be applied to cross-sectional hierarchies

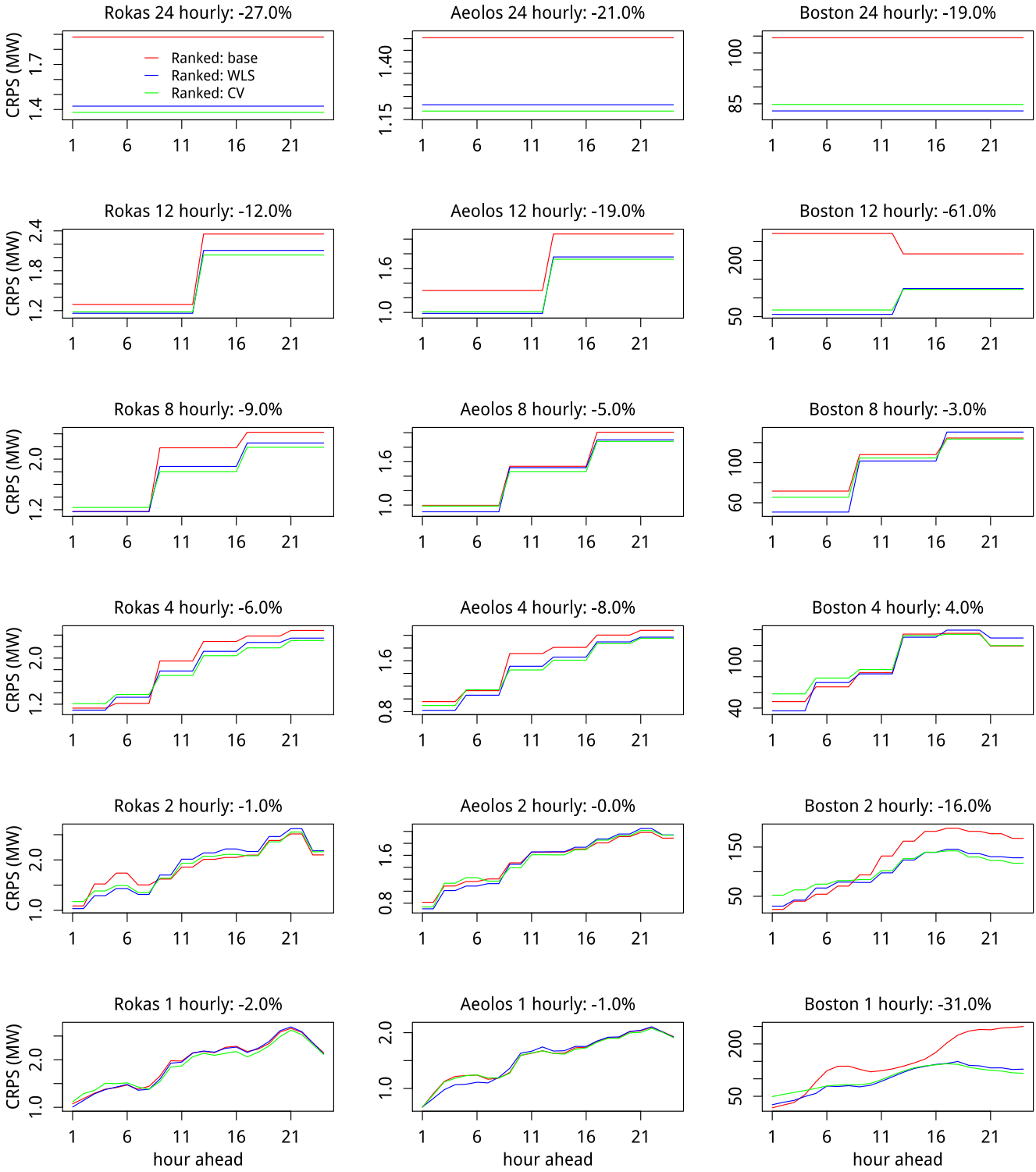

Figure 11: Probabilistic Evaluation of wind power forecasts in the evaluation period using CRPS for the Rokas and Aeolos wind farms and the electric load of Boston, comparing (1) base using the ranked sample and (2) WLS using the ranked sample and (3) Cross-validated using the ranked sample with sum-to-one weights. The improvement of cross-validated over base is presented in average percentage for each level separately, on top of each plot. Lower values are better.

with little to no loss of generality. We proposed three schemes for obtaining samples of base forecasts from the joint predictive densities, namely permuted, ranked and stacked sampling. These approaches correspond to the cases of no dependence between the hierarchical nodes, comonotonic dependence between nodes and temporal model driven dependencies within a level respectively. Since these samples do not respect aggregation constraints they are subsequently reconciled using the bottom-up, global average and WLS approaches. Furthermore, we investigated for the first time the use of a cross-validation approach for obtaining the reconciliation weights. The performance of the various combinations of sampling schemes and reconciliation methods was subsequently measured by producing and evaluating probabilistic wind power and electric load forecasts.

The empirical results from two wind farms in Greece and from energy load in Boston suggest that cross-validation reconciliation based on ranked samples offers the best performance overall compared to all other approaches. Performance enhancement for density forecast evaluation in terms of energy score is up to 15% for wind power data and up to 40% for load data relative to base. For wind data, lower resolutions (higher levels of aggregation) enjoyed the most performance benefits, providing direct managerial benefits for transmission operations and planning for optimal trading strategies. The results also show that comonotonic aggregation of quantiles is the most robust method when using reconciliation.

Looking forward, our research also poses new research questions that lie outside the scope of the current paper. For example, although an advantage of the stacked sample, ranked sample and permuted sample is their ease of construction, it may be worthwhile developing more complicated merging schemes. These could be based on the algorithms developed by Ben Taieb, Taylor, and Hyndman (2017) thus extending that method to utilise information about the entire predictive density of the full hierarchy of base forecasts. It may also be worthwhile investigating whether the sparse structure of the $\boldsymbol{P}$ matrix can be selected in a more data driven way, especially for cross sectional hierarchies where a different pattern of sparsity may be required. Finally, it would be interesting to see if methods that can generate density forecasts based on ensembles or physical models can also benefit from using (temporal) hierarchical reconciliation.

Acknowledgements

Jooyoung Jeon was supported by the EPSRC grant (EP/N03466X/1). We are grateful to George Sideratos of the National Technical University of Athens and the EU SafeWind Project for providing the data. We are also grateful for the insightful comments of participants at the International Symposium on Energy Analytics in Cairns, Australia, 2017 and to two anonymous referees.

References

Abouarghoub, Wessam, Nikos K Nomikos, and Fotios Petropoulos. 2018. "On reconciling macro and micro energy transport forecasts for strategic decision making in the tanker industry." *Transportation Research Part E: Logistics and Transportation Review* 113: 225–238.

Arbenz, Philipp, Christoph Hummel, and Georg Mainik. 2012. "Copula based hierarchical risk aggregation through sample reordering." *Insurance: Mathematics and Economics* 51 (1): 122–133.

Athanasopoulos, George, Roman A Ahmed, and Rob J Hyndman. 2009. "Hierarchical forecasts for Australian domestic tourism." *International Journal of Forecasting* 25 (1): 146–166.

Athanasopoulos, George, Rob J. Hyndman, Nikolaos Kourentzes, and Fotios Petropoulos. 2017. "Forecasting with temporal hierarchies." *European Journal of Operational Research* 262 (1): 60–74.

Baillie, Richard T., Tim Bollerslev, and Hans Ole Mikkelsen. 1996. "Fractionally integrated generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics* 74 (1): 3–30.

Ben Taieb, Souhaib, James W. Taylor, and Rob J. Hyndman. 2017. "Coherent probabilistic forecasts for hierarchical time series." *Proceedings of the 34th International Conference on Machine Learning* 70: 3348–3357.

Bollerslev, Tim, Robert F. Engle, and Jeffrey M. Wooldridge. 1988. "A capital asset pricing model with time-varying covariances." *Journal of Political Economy* 96 (1): 116–131.

Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society: Series B* 26 (2): 211–243.

Dangerfield, Byron J, and John S Morris. 1992. "Top-down or bottom-up: Aggregate versus disaggregate extrapolations." *International Journal of Forecasting* 8 (2): 233–241.

Dowell, Jethro, and Pierre Pinson. 2015. "Very-short-term probabilistic wind power forecasts by sparse vector autoregression." *IEEE Transactions on Smart Grid* 7 (2): 763–770.

Fan, Shu, and Rob J. Hyndman. 2011. "Short-Term Load Forecasting Based on a Semi-Parametric Additive Model." *IEEE Transactions on Power Systems* 27 (1): 134–141.

Fliedner, Gene. 1999. "An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation." *Computers & Operations Research* 26 (10): 1133–1149.

Gamakumara, Puwasala, Anastasios Panagiotelis, George Athanasopoulos, Rob J Hyndman, et al. 2018. "Probabilistic Forecasts in Hierarchical Time Series." *Monash University: Melbourne, Australia* .

Gneiting, Tilmann, and Matthias Katzfuss. 2014. "Probabilistic forecasting." *Annual Review of Statistics and Its Application* 1 (1): 125–151.

Gneiting, Tilmann, Kristin Larson, Kenneth Westrick, Marc G Genton, and Eric Aldrich. 2006. "Calibrated probabilistic forecasting at the stateline wind energy center." *Journal of the American Statistical Association* 101 (475): 968–979.

Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378.

Granger, C. W. J., and Roselyne Joyeux. 1980. "An introduction to long-memory time series models and fractional differencing." *Journal of Time Series Analysis* 1 (1): 15–29.

Gross, Charles W, and Jeffrey E Sohl. 1990. "Disaggregation methods to expedite product line forecasting." *Journal of Forecasting* 9 (3): 233–254.

Hering, Amanda S., and Marc G. Genton. 2010. "Powering up with space-time wind forecasting." *Journal of the American Statistical Association* 105 (489): 92–104.

Hong, Tao, and Shu Fan. 2016. "Probabilistic electric load forecasting: A tutorial review." *International Journal of Forecasting* 32 (3): 914–938.

Hong, T., J. Xie, and J. Black. In press. "Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting." *International Journal of Forecasting* .

Hosking, J. R. M. 1981. "Fractional differencing." *Biometrika* 68 (1): 165.

Hyndman, Athanasopoulos G. Razbash S. Schmidt D. Zhou Z. Khan Y. Wang E., R. J. 2014. "R Package: forecast." http://cran.r-project.org/web/packages/forecast/index.html.

Hyndman, Rob J., Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang. 2011. "Optimal combination forecasts for hierarchical time series." *Computational Statistics & Data Analysis* 55 (9): 2579–2589.

Hyndman, Rob J, and George Athanasopoulos. 2018. *Forecasting: principles and practice*. OTexts.

Hyndman, Rob J., and Shu Fan. 2010. "Density forecasting for long-term peak electricity demand." *IEEE Transactions on Power Systems* 25 (2): 1142–1153,.

Hyndman, Rob J, Alan J Lee, and Earo Wang. 2016. "Fast computation of reconciled forecasts for hierarchical and grouped time series." *Computational Statistics & Data Analysis* 97: 16–32.

Jeon, Jooyoung, and James W. Taylor. 2012. "Using conditional kernel density estimation for wind power density forecasting." *Journal of the American Statistical Association* 107 (497): 66–79.

Koning, Alex J, Philip Hans Franses, Michele Hibon, and Herman O Stekler. 2005. "The M3 competition: Statistical tests of the results." *International Journal of Forecasting* 21 (3): 397–409.

Kourentzes, Nikolaos, and Fotios Petropoulos. 2016. "Forecasting with multivariate temporal aggregation: The case of promotional modelling." *International Journal of Production Economics* 181, Part A: 145–153.

Kourentzes, Nikolaos, Fotios Petropoulos, and Juan Ramon Trapero. 2014. "Improving forecasting by estimating time series structural components across multiple frequencies." *International Journal of Forecasting* 30 (2): 291–302.

Lichtendahl, Kenneth C., Yael Grushka-Cockayne, and Robert L. Winkler. 2013. "Is it better to average probabilities or quantiles?." *Management Science* 59 (7): 1594–1611.

Lütkepohl, Helmut. 1984. "Forecasting contemporaneously aggregated vector ARMA processes." *Journal of Business & Economic Statistics* 2 (3): 201–214.

Morstyn, Thomas, Niall Farrell, Sarah J. Darby, and Malcolm D. McCulloch. 2018. "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants." *Nature Energy* 3: 94–101.

Nikolopoulos, Konstantinos, Aris A Syntetos, John E Boylan, Fotios Petropoulos, and Vassilios Assimakopoulos. 2011. "An aggregate - disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis." *Journal of the Operational Research Society* 62 (3): 544–554.

Nystrup, Peter, Erik Lindström, Pierre Pinson, and Henrik Madsen. 2018. "Temporal hierarchies with autocorrelation for load forecasting." *Working Paper* `http://pierrepinson.com/docs/Nystrupetal2019.pdf`.

Petropoulos, Fotios, and Nikolaos Kourentzes. 2014. "Improving forecasting via multiple temporal aggregation." *Foresight: The International Journal of Applied Forecasting* 34: 12–17.

Petropoulos, Fotios, and Nikolaos Kourentzes. 2015. "Forecast combinations for intermittent demand." *The Journal of the Operational Research Society* 66 (6): 914–924.

Petropoulos, Fotios, Nikolaos Kourentzes, and Konstantinos Nikolopoulos. 2016. "Another look at estimators for intermittent demand." *International Journal of Production Economics* 181, Part A: 154–161.

Pinson, Pierre. 2013. "Wind energy: forecasting challenges for its operational management." *Statistical Science* 28 (4): 564–585.

Roach, C. In press. "Reconciled boosted models for GEFCom2017 hierarchical probabilistic load forecasting." *International Journal of Forecasting* .

Rostami-Tabar, B., M.Z. Babai, A. Syntetos, and Y. Ducq. 2013. "Demand forecasting by temporal aggregation." *Naval Research Logistics* 60 (6).

Roulston, M.S., and L.A. Smith. 2003. "Combining dynamical and statistical ensembles." *Tellus A: Dynamic Meteorology and Oceanography* 55 (1): 16–30.

Sloughter, J. McLean, Tilmann Gneiting, and Adrian E. Raftery. 2010. "Probabilistic wind speed forecasting using ensembles and Bayesian model averaging." *Journal of the American Statistical Association* 105 (489): 25–35.

Spiliotis, Evangelos, Fotios Petropoulos, Nikolaos Kourentzes, and Vassilios Assimakopoulos. 2018. "Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption." *Forecasting and Strategy Unit Working Paper 1/18* `http://www.fsu.gr/en/research/working-papers`.

Spithourakis, Georgios, Fotios Petropoulos, Konstantinos Nikolopoulos, and Vassilios Assimakopoulos. 2014. "A systemic view of ADIDA framework." *IMA Journal of Management Mathematics* 25 (2): 125–137.

Taylor, J.W., P.E. McSharry, and R. Buizza. 2009. "Wind power density forecasting using ensemble predictions and time series models." *IEEE Transactions on Energy Conversion* 24: 775–782.

Taylor, J W. 2003. "Short-Term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing." *The Journal of the Operational Research Society* 54 (8): 799–805.

Taylor, James W. 2017. "Probabilistic forecasting of wind power ramp events using autoregressive logit models." *European Journal of Operational Research* 259 (2): 703–712.

Taylor, James W., and Jooyoung Jeon. 2015. "Forecasting wind power quantiles using conditional kernel estimation." *Renewable Energy* 80: 370–379.

Wickramasuriya, Shanika L, George Athanasopoulos, and Rob J Hyndman. 2018. "Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization." *Journal of the American Statistical Association* 1–16.

Zellner, Arnold, and Justin Tobias. 2000. "A note on aggregation, disaggregation and forecasting performance." *Journal of Forecasting* 19 (5): 457–465.